

AN OPTIMAL AFFINE INVARIANT SMOOTH MINIMIZATION ALGORITHM

ALEXANDRE D'ASPREMONT AND MARTIN JAGGI

ABSTRACT. We formulate an affine invariant implementation of the algorithm in [Nesterov, 1983]. We show that the complexity bound is then proportional to an affine invariant regularity constant defined with respect to the Minkowski gauge of the feasible set.

1. INTRODUCTION

In this short note, we show how to implement the smooth minimization algorithm described in [Nesterov, 1983, 2005] so that both its iterations and its complexity bound are invariant by a change of coordinates in the problem. We focus on the minimization problem

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && x \in Q, \end{aligned} \tag{1}$$

where f is a convex function with Lipschitz continuous gradient and Q is a compact convex set. Without too much loss of generality, we will assume that the interior of Q is nonempty and contains zero. When Q is sufficiently simple, in a sense that will be made precise later, Nesterov [1983] showed that this problem could be solved with a complexity of $O(1/\sqrt{\epsilon})$, where ϵ is the precision target. Furthermore, it can be shown that this complexity bound is optimal for the class of smooth problems [Nesterov, 2003].

While the dependence in $O(1/\sqrt{\epsilon})$ of the complexity bound in Nesterov [1983] is optimal, the constant in front of that bound still depends on a few parameters which vary with implementation: the choice of norm and prox regularization function. This means in particular that, everything else being equal, this bound is not invariant with respect to an affine change of coordinates, so the complexity bound varies while the intrinsic complexity of problem (1) remains unchanged. Here, we show one possible fix for this inconsistency, by choosing a norm and a prox term for the algorithm in [Nesterov, 1983, 2005] which make its iterations and complexity invariant by a change of coordinates.

2. SMOOTH OPTIMIZATION ALGORITHM

We first recall the basic structure of the algorithm in [Nesterov, 1983]. While many variants of this method have been derived, we use the formulation in [Nesterov, 2005]. We choose a norm $\|\cdot\|$ and assume that the function f in problem (1) is convex with Lipschitz continuous gradient, so

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2, \quad x, y \in Q, \tag{2}$$

for some $L > 0$. We also choose a prox function $d(x)$ for the set Q , i.e. a continuous, strongly convex function on Q with parameter σ (see Nesterov [2003] or Hiriart-Urruty and Lemaréchal [1993] for a discussion of regularization techniques using strongly convex functions). We let x_0 be the center of Q for the prox-function $d(x)$ so that

$$x_0 \triangleq \operatorname{argmin}_{x \in Q} d(x),$$

assuming w.l.o.g. that $d(x_0) = 0$, we then get in particular

$$d(x) \geq \frac{1}{2}\sigma\|x - x_0\|^2. \tag{3}$$

We write $T_Q(x)$ a solution to the following subproblem

$$T_Q(x) \triangleq \operatorname{argmin}_{y \in Q} \left\{ \langle \nabla f(x), y - x \rangle + \frac{1}{2} L \|y - x\|^2 \right\} \quad (4)$$

We let $y_0 \triangleq T_Q(x_0)$ where x_0 is defined above. We recursively define three sequences of points: the current iterate x_k , the corresponding $y_k = T_Q(x_k)$, and the points

$$z_k \triangleq \operatorname{argmin}_{x \in Q} \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \right\} \quad (5)$$

and a step size sequence $\alpha_k \geq 0$ with $\alpha_0 \in (0, 1]$ so that

$$\begin{aligned} x_{k+1} &= \tau_k z_k + (1 - \tau_k) y_k \\ y_{k+1} &= T_Q(x_{k+1}) \end{aligned} \quad (6)$$

where $\tau_k = \alpha_{k+1}/A_{k+1}$ with $A_k = \sum_{i=0}^k \alpha_i$. We implicitly assume here that Q is simple enough so that the two subproblems defining y_k and z_k can be solved very efficiently. We have the following convergence result.

Theorem 2.1. *Suppose $\alpha_k = (k + 1)/2$ with the iterates x_k , y_k and z_k defined in (5) and (6), then for any $k \geq 0$ we have*

$$f(y_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)^2}$$

where x^* is an optimal solution to problem (1).

Proof. See [Nesterov \[2005\]](#). ■

If $\epsilon > 0$ is the target precision, Theorem 2.1 ensures that Algorithm 1 will converge to an ϵ -accurate solution in no more than

$$\sqrt{\frac{8Ld(x^*)}{\sigma\epsilon}} \quad (7)$$

iterations. In practice of course, $d(x^*)$ needs to be bounded a priori and L and σ are often hard to evaluate.

Algorithm 1 Smooth minimization.

Input: x_0 , the prox center of the set Q .

- 1: **for** $k = 0, \dots, N$ **do**
- 2: Compute $\nabla f(x_k)$.
- 3: Compute $y_k = T_Q(x_k)$.
- 4: Compute $z_k = \operatorname{argmin}_{x \in Q} \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^k \alpha_i [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \right\}$.
- 5: Set $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$.
- 6: **end for**

Output: $x_N, y_N \in Q$.

While most of the parameters in Algorithm 1 are set explicitly, the norm $\|\cdot\|$ and the prox function $d(x)$ are chosen arbitrarily. In what follows, we will see that a natural choice for both makes the algorithm affine invariant.

3. AFFINE INVARIANT IMPLEMENTATION

We can define an affine change of coordinates $x = Ay$ where $A \in \mathbb{R}^{n \times n}$ is a nonsingular matrix, for which the original optimization problem in (1) is transformed so

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array} \quad \text{becomes} \quad \begin{array}{ll} \text{minimize} & \hat{f}(y) \\ \text{subject to} & y \in \hat{Q}, \end{array} \quad (8)$$

in the variable $y \in \mathbb{R}^n$, where

$$\hat{f}(y) \triangleq f(Ay) \quad \text{and} \quad \hat{Q} \triangleq A^{-1}Q. \quad (9)$$

Unless A is pathologically ill-conditioned, both problems are equivalent and should have invariant complexity bounds and iterations. In fact, the complexity analysis of Newton's method based on the self-concordance argument developed in [Nesterov and Nemirovskii, 1994] produces affine invariant complexity bounds and the iterates themselves are invariant. Here we will show how to choose the norm $\|\cdot\|$ and the prox function $d(x)$ to get a similar behavior for Algorithm 1.

3.1. Choosing the norm. We start by a few classical results and definitions. Recall that the *Minkowski gauge* of a set Q is defined as follows.

Definition 3.1. Let $Q \subset \mathbb{R}^n$ containing zero, we define the *Minkowski gauge* of Q as

$$\gamma_Q(x) \triangleq \inf\{\lambda \geq 0 : x \in \lambda Q\}$$

with $\gamma_Q(x) = 0$ when Q is unbounded in the direction x .

When Q is a compact convex, centrally symmetric set with respect to the origin and has nonempty interior, the Minkowski gauge defines a *norm*. We write this norm $\|\cdot\|_Q \triangleq \gamma_Q(\cdot)$. From now on, we will assume that the set Q is centrally symmetric or use for example $\bar{Q} = Q - Q$ (in the Minkowski sense) for the gauge when it is not (this can be improved and extending these results to the nonsymmetric case is a classical topic in functional analysis). Note that any affine transform of a centrally symmetric convex set remains centrally symmetric. The following simple result shows why $\|\cdot\|_Q$ is potentially a good choice of norm for Algorithm 1.

Lemma 3.2. Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$, Q is a centrally symmetric convex set with nonempty interior and let $A \in \mathbb{R}^{n \times n}$ be a nonsingular matrix. Then f has Lipschitz continuous gradient with respect to the norm $\|\cdot\|_Q$ with constant $L > 0$, i.e.

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|_Q^2, \quad x, y \in Q,$$

if and only if the function $f(Aw)$ has Lipschitz continuous gradient with respect to the norm $\|\cdot\|_{A^{-1}Q}$ with the same constant L .

Proof. Let $y = Az$ and $x = Aw$, then

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|_Q^2, \quad x, y \in Q,$$

is equivalent to

$$f(Az) \leq f(Aw) + \langle A^{-T} \nabla_w f(Aw), Az - Aw \rangle + \frac{1}{2}L\|Az - Aw\|_Q^2, \quad z, w \in A^{-1}Q,$$

and, using the fact that $\|Ax\|_Q = \|x\|_{A^{-1}Q}$, this is also

$$f(Az) \leq f(Aw) + \langle \nabla_w f(Aw), A^{-1}(Az - Aw) \rangle + \frac{1}{2}L\|z - w\|_{A^{-1}Q}^2, \quad z, w \in A^{-1}Q,$$

hence the desired result. ■

An almost identical argument shows the following analogous result for the property of *strong convexity* with respect to the norm $\|\cdot\|_Q$ and affine changes of coordinates.

Lemma 3.3. Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$, Q is a centrally symmetric convex set with nonempty interior and let $A \in \mathbb{R}^{n \times n}$ be a nonsingular matrix. Suppose f is strongly convex with respect to the norm $\|\cdot\|_Q$ with parameter $\sigma > 0$, i.e.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\sigma \|y - x\|_Q^2, \quad x, y \in Q,$$

if and only if the function $f(Ax)$ is strongly convex with respect to the norm $\|\cdot\|_{A^{-1}Q}$ with the same parameter σ .

We now turn our attention to the choice of prox function in Algorithm 1.

3.2. Choosing the prox. Choosing the norm as $\|\cdot\|_Q$ allows us to define a norm without introducing an arbitrary geometry in the algorithm, since the norm is extracted directly from the problem definition. When Q is smooth, a similar reasoning allows us to choose the prox term in Algorithm 1, and we can set $d(x) = \|x\|_Q^2$. The immediate impact of this choice is that the term $d(x^*)$ in (7) is bounded by one, by construction. This choice has other natural benefits which are highlighted below. We first recall a result showing that the conjugate of a squared norm is the squared dual norm.

Lemma 3.4. Let $\|\cdot\|$ be a norm and $\|\cdot\|^*$ its dual norm, then

$$\frac{1}{2} (\|y\|^*)^2 = \sup_x y^T x - \frac{1}{2} \|x\|^2.$$

Proof. We recall the proof in [Boyd and Vandenberghe, 2004, Example 3.27] as it will prove useful in what follows. By definition, $x^T y \leq \|y\|^* \|x\|$, hence

$$y^T x - \frac{1}{2} \|x\|^2 \leq \|y\|^* \|x\| - \frac{1}{2} \|x\|^2 \leq \frac{1}{2} (\|y\|^*)^2$$

because the second term is a quadratic function of $\|x\|^2$, with maximum $(\|y\|^*)^2/2$. This maximum is attained by any x such that $x^T y = \|y\|^* \|x\|$ (there must be one by construction of the dual norm), normalized so $\|x\| = \|y\|^*$, which yields the desired result. ■

This last result (and its proof) shows that solving the prox mapping is equivalent to finding a vector aligned with the gradient, with respect to the Minkowski norm $\|\cdot\|_Q$. We now recall another simple result showing that the dual of the norm $\|\cdot\|_Q$ is given by $\|\cdot\|_{Q^\circ}$ where Q° is the polar of Q .

Lemma 3.5. Let Q be a centrally symmetric convex set with nonempty interior, then $\|\cdot\|_Q^* = \|\cdot\|_{Q^\circ}$.

Proof. We write

$$\begin{aligned} \|x\|_{Q^\circ} &= \inf\{\lambda \geq 0 : x \in \lambda Q^\circ\} \\ &= \inf\{\lambda \geq 0 : x^T y \leq \lambda, \text{ for all } y \in Q\} \\ &= \inf\left\{\lambda \geq 0 : \sup_{y \in Q} x^T y \leq \lambda\right\} \\ &= \sup_{y \in Q} x^T y \\ &= \|x\|_Q^* \end{aligned}$$

which is the desired result. ■

The last remaining issue to settle is the strong convexity of the squared Minkowski norm. Fortunately, this too is a classical result in functional analysis, as a squared norm is strongly convex with respect to itself if and only if its dual norm has a smoothness modulus of power 2.

However, this does not cover the case where the norm $\|\cdot\|_Q$ is not smooth. In that scenario, we need to pick the norm based on Q but find a smooth prox function not too different from $\|\cdot\|_Q$. This is exactly the problem studied by Juditsky and Nemirovski [2008] who define the regularity of a Banach space $(E, \|\cdot\|_E)$

in terms of the smoothness of the best smooth approximation of the norm $\|\cdot\|_E$. We first recall a few more definitions, and we will then show that the regularity constant defined by [Juditsky and Nemirovski \[2008\]](#) produces an affine invariant bound on the term $d(x^*)/\sigma$ in the complexity of the smooth algorithm in [\[Nesterov, 1983\]](#).

Definition 3.6. Suppose $\|\cdot\|_X$ and $\|\cdot\|_Y$ are two norms on a space E , the distortion $d(\|\cdot\|_X, \|\cdot\|_Y)$ between these two norms is equal to the smallest product $ab > 0$ such that

$$\frac{1}{b}\|x\|_Y \leq \|x\|_X \leq a\|x\|_Y$$

over all $x \in E$.

Note that $\log d(\|\cdot\|_X, \|\cdot\|_Y)$ defines a metric on the set of all symmetric convex bodies in \mathbb{R}^n , called the *Banach-Mazur distance*. We then recall the regularity definition in [Juditsky and Nemirovski \[2008\]](#).

Definition 3.7. The regularity constant of a Banach space $(E, \|\cdot\|)$ is the smallest constant $\Delta > 0$ for which there exists a smooth norm $p(x)$ such that

- (i) $p(x)^2/2$ has a Lipschitz continuous gradient with constant μ w.r.t. the norm $p(x)$, with $1 \leq \mu \leq \Delta$,
- (ii) the norm $p(x)$ satisfies

$$\|x\|^2 \leq p(x)^2 \leq \frac{\Delta}{\mu} \|x\|^2, \quad \text{for all } x \in E \quad (10)$$

$$\text{hence } d(p(x), \|\cdot\|) \leq \sqrt{\Delta/\mu}.$$

Note that in finite dimension, since all norms are equivalent to the Euclidean norm with distortion at most $\sqrt{\dim E}$, we know that all finite dimensional Banach spaces are at least $(\dim E)$ -regular. Furthermore, the regularity constant is invariant with respect to an affine change of coordinates since both the distortion and the smoothness bounds are. We are now ready to prove the main result of this section.

Proposition 3.8. Let $\epsilon > 0$ be the target precision, suppose that the function f has a Lipschitz continuous gradient with constant L_Q with respect to the norm $\|\cdot\|_Q$ and that the space $(\mathbb{R}^n, \|\cdot\|_Q^*)$ is D_Q -regular, then Algorithm 1 will produce an ϵ -solution to problem (1) in at most

$$\sqrt{8 \frac{L_Q \min\{D_Q/2, n\}}{\epsilon}} \quad (11)$$

iterations. The constants L_Q and D_Q are affine invariant.

Proof. If $(\mathbb{R}^n, \|\cdot\|_Q^*)$ is D_Q -regular, then by Definition 3.7, there exists a norm $p(x)$ such that $p(x)^2/2$ has a Lipschitz continuous gradient with constant μ with respect to the norm $p(x)$, and [\[Juditsky and Nemirovski, 2008, Prop. 3.2\]](#) shows by conjugacy that the prox function $d(x) \triangleq p^*(x)^2/2$ is strongly convex with respect to the norm $p^*(x)$ with constant $1/\mu$. Now (10) means that

$$\sqrt{\frac{\mu}{D_Q}} \|x\|_Q \leq p^*(x) \leq \|x\|_Q, \quad \text{for all } x \in Q$$

since $\|\cdot\|^{**} = \|\cdot\|$, hence

$$\begin{aligned} d(x+y) &\geq d(x) + \langle \partial d(x), y \rangle + \frac{1}{2\mu} p^*(y)^2 \\ &\geq d(x) + \langle \partial d(x), y \rangle + \frac{1}{2D_Q} \|y\|_Q^2 \end{aligned}$$

so $d(x)$ is strongly convex with respect to $\|\cdot\|_Q$ with constant $\sigma = 1/D_Q$, and using (10) as above

$$\frac{d(x^*)}{\sigma} = \frac{p^*(x^*)^2 D_Q}{2} \leq \frac{\|x^*\|_Q^2 D_Q}{2} \leq \frac{D_Q}{2}$$

by definition of $\|\cdot\|_Q$, if x^* is an optimal (hence feasible) solution of problem (1). The bound in (11) then follows from (7) and its affine invariance follows directly from affine invariance of the distortion and Lemmas 3.2 and 3.3. ■

4. EXAMPLES

To illustrate our results, consider the problem of minimizing a smooth convex function over the unit simplex, written

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && \mathbf{1}^T x \leq 1, x \geq 0, \end{aligned} \tag{12}$$

in the variable $x \in \mathbb{R}^n$. As discussed in [Juditsky et al., 2009, §3.3], choosing $\|\cdot\|_1$ as the norm and $d(x) = \log n + \sum_{i=1}^n x_i \log x_i$ as the prox function, we have $\sigma = 1$ and $d(x^*) \leq \log n$, which means the complexity of solving (12) using Algorithm 1 is bounded by

$$8\sqrt{\frac{L_1 \log n}{\epsilon}} \tag{13}$$

where L_1 is the Lipschitz constant of ∇f with respect to the ℓ_1 norm. This choice of norm and prox has a double advantage here. First, the prox term $d(x^*)$ grows only as $\log n$ with the dimension. Second, the ℓ_∞ norm being the smallest among all ℓ_p norms, the smoothness bound L_1 is also minimal among all choices of ℓ_p norms.

Let us now follow the construction of Section 3. The simplex $C = \{x \in \mathbb{R}^n : \mathbf{1}^T x \leq 1, x \geq 0\}$ is not centrally symmetric, but we can symmetrize it as the ℓ_1 ball. The Minkowski norm associated with that set is then equal to the ℓ_1 -norm, so $\|\cdot\|_Q = \|\cdot\|_1$ here. The space $(\mathbb{R}^n, \|\cdot\|_\infty)$ is $2 \log n$ regular [Juditsky and Nemirovski, 2008, Example 3.2] with the prox function chosen here as $\|\cdot\|_{(2 \log n)}^2/2$. Proposition 3.8 then shows that the complexity bound we obtain using this procedure is identical to that in (13). A similar result holds in the matrix case.

5. CONCLUSION

From a practical point of view, the results above offer guidance in the choice of a prox. function depending on the geometry of the feasible set Q . On the theoretical side, these results provide affine invariant descriptions of the complexity of the feasible set and of the smoothness of the objective function, written in terms of the regularity constant of the polar of the feasible set and the Lipschitz constant of ∇f with respect to the Minkowski norm. However, while we show that it is possible to formulate an affine invariant implementation of the optimal algorithm in [Nesterov, 1983], we do not yet show that this is always a good idea... In particular, given our choice of norm the constants L_Q and D_Q are both affine invariant, with L_Q optimal by construction and our choice of prox function minimizing D_Q over all smooth square norms, but this does not mean that our choice of norm (Minkowski) minimizes the product $L_Q \min\{D_Q/2, n\}$, hence that we achieve the best possible bound for the complexity of the smooth algorithm in [Nesterov, 1983].

ACKNOWLEDGMENTS

AA and MJ would like to acknowledge support from the European Research Council (project SIPA). MJ also acknowledges support from the Swiss National Science Foundation (SNSF).

REFERENCES

- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer, 1993.
- A. Juditsky and A.S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.
- A. Juditsky, G. Lan, A. Nemirovski, and A. Shapiro. Stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

- Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.
- Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2003.
- Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

CMAP, UMR CNRS 7641, ECOLE POLYTECHNIQUE, PALAISEAU, FRANCE.
E-mail address: alexandre.daspremont@m4x.org

CMAP, UMR CNRS 7641, ECOLE POLYTECHNIQUE, PALAISEAU, FRANCE.
E-mail address: martin.jaggi@polytechnique.fr